Alaina Rongione

DATA 400: Capstone

Professor Bilen

11 September 2025

Project Idea

The three course sequence chosen to complete the Data Analytics major was English, so the aim for the project is to apply data-driven methods to examine the receptions of classic literature and their movie adaptations. The proposed research question is: "How does audience sentiment differ between classic literature and their movies, and what common factors impact these differences?" Ultimately, the results of the research should conclude whether the books or their movie adaptations are viewed as "better" in the eyes of the public (this can be concluded in ratings). Furthermore, examining the feedback from the readers and viewers will highlight the themes and qualities for which either the books or movies excel.

The data will be retrieved from two websites: Goodreads (for the classic novels) and IMDB (for the movie adaptations). Both websites contain ratings and comments from the audiences for the books and movies. The summaries of each novel and movie will also be scraped for context and possible discoveries. Once the data is collected and because it is a comparison question, inferential data models will mainly be utilized during the course of the project. First, finding the average rating of each book and movie will be necessary to begin. Pairing a visualization such as histograms to compare the average ratings of classic novels and movies will be helpful in understanding the rating distributions. The intention is to also find the common themes and comments between both the books and the movies. From there, the differences will be found through topic modeling and visualizations through word bubbles. If any large themes or characteristics appear in the book's comments or the movie's comments. I will do deeper analysis on the comments to find if those differences correlate to the rating differences.

Overall, the conclusions of this project could influence decision-making for authors and producers in writing books or creating film adaptations (the project informs them on what the audience cares about) and could influence marketing strategies (where promoting in a certain manner that attracts more people). Additionally, literature and movie lovers will benefit from the new actions of the writers and producers in receiving books/movies they enjoy more. Ultimately, the implications of the stakeholders are mainly making decisions based on data that enhances success and receptions of books and films.

Socially, the project sheds new light on how adaptations are reshaped and form new opinions among the audience on historical and cultural narratives. Additionally, these data results can drive renewed interest in classic literature that has been lost over time. Finally, another societal implication is that the project could possibly reveal new opinions about inclusivity and representation in media, discussing gender roles, race, and cultural diversity. Legally, check to see if Goodreads and IMDB offer cheap and accessible APIs to gather the data. However, I will need to double check the Terms of Use to see if scraping data is legal for educational purposes. If not, I must travel to different, trustworthy websites with ratings. I know Goodreads has an API to use, but IMDB may be different. The secondary source if IMDB falls through is Rotten Tomatoes. Finally, in terms of ethics, I will need to be careful to respect data privacy among the commenters and represent their opinions in a respectful and accurate manner. In other words, I cannot let personal bias and judgement skew results, and I must be careful to not misinterpret or construe the audience's comments.